Problem 1 (Expectation-Maximization for Categorical-Geometric Mixture Models, 25pts)

In this problem we will explore expectation-maximization for a Categorical-Geometric Mixture model.

Specifically, we assume that there are a set of K parameters $p_k \in [0, 1]$. To generate each observation n, we first choose which of those K parameters we will use based on the (unknown) overall mixing proportion over the components $\boldsymbol{\theta} \in [0, 1]^K$, where $\sum_{k=1}^{K} \theta_k = 1$. Let the (latent) $\mathbf{z}_n$ indicate which of the K components we use to generate observation n. Next we sample the observation x_n from a geometric distribution with parameter $p_{\mathbf{z}_n}$. This process can be written as:

$$\begin{aligned} \mathbf{z}_n &\sim \text{Categorical}(\boldsymbol{\theta}) \\ x_n &\sim \text{Geometric}(p_{\mathbf{z}_n}) \end{aligned}$$

We encode observation n's latent component-assignment $\mathbf{z}_n \in \{0, 1\}^K$ as a one-hot vector. Element indicator variables z_{nk} equal 1 if $\mathbf{z}_n$ was generated using component k.

A geometric distribution corresponds to the number of trials needed to get to the first success. If success occurs with probability p, its PMF is given by $p(x_n; p_k) = (1 - p_k)^{x_n - 1} p_k$, where $x_n \in \{1, 2, ...\}$.

1. **Intractability of the Data Likelihood** We are generally interested in finding a set of parameters p_k that maximize the data likelihood $\log p(\{x_n\}_{n=1}^N; \boldsymbol{\theta}, \{p_k\}_{k=1}^K)$. Expand the data likelihood to include the necessary sums over observations x_n and to marginalize out (via more sums) the latents $\mathbf{z}_n$. Why is optimizing this likelihood directly intractable?

2. **Complete-Data Log Likelihood** The complete data $D = \{(x_n, \mathbf{z}_n)\}_{n=1}^N$ includes latents $\mathbf{z}_n$. Write out the complete-data negative log likelihood. Apply the "power trick" and simplify your expression using indicator elements z_{nk}. [a]

   $$\mathcal{L}(\boldsymbol{\theta}, \{p_k\}_{k=1}^K) = -\ln p(D; \boldsymbol{\theta}, \{p_k\}_{k=1}^K).$$

   Note that optimizing this loss is now computationally tractable if we know $\mathbf{z}_n$.

3. **Expectation Step** Our next step is to introduce a mathematical expression for $\mathbf{q}_n$, the posterior over the hidden component variables $\mathbf{z}_n$ conditioned on the observed data x_n with fixed parameters. That is:

   $$\mathbf{q}_n = \begin{bmatrix} p(\mathbf{z}_n = \mathbf{C}_1 | x_n; \boldsymbol{\theta}, \{p_k\}_{k=1}^K) \\ \cdots \\ \cdots \\ p(\mathbf{z}_n = \mathbf{C}_K | x_n; \boldsymbol{\theta}, \{p_k\}_{k=1}^K) \end{bmatrix}$$

   where $\mathbf{C}_k$ is a 1-hot encoded vector to represent component k.

   - **Part 3.A** Write down and simplify the expression for $\mathbf{q}_n$. Note that because the $\mathbf{q}_n$ represents the posterior over the hidden categorical variables $\mathbf{z}_n$, the components of vector $\mathbf{q}_n$ must sum to 1. The main work is to find an expression for $p(\mathbf{z}_n | x_n; \boldsymbol{\theta}, \{p_k\}_{k=1}^K)$ for any choice of $\mathbf{z}_n$; i.e., for any 1-hot encoded $\mathbf{z}_n$. With this, you can then construct the different components that make up the vector $\mathbf{q}_n$.
   - **Part 3.B** Give a concrete algorithm for calculating $\mathbf{q}_n$ for each example x_n, given parameters $\boldsymbol{\theta}$ and $\{p_k\}_{k=1}^K$.

(Continued on next page.)

[a] The "power trick" is used when terms in a PDF are raised to the power of indicator components of a one-hot vector. For example, it allows us to rewrite $p(\mathbf{z}_n; \boldsymbol{\theta}) = \prod_k \theta_k^{z_{nk}}$.

Problem 1 (cont.)

4. **Maximization Step** Using the $\mathbf{q}_n$ estimates from the Expectation Step, derive an update for maximizing the expected complete data log likelihood in terms of $\boldsymbol{\theta}$ and $\{p_k\}_{k=1}^K$.

   - **Part 4.A** Derive an expression for the expected complete-data log likelihood using $\mathbf{q}_n$.
   - **Part 4.B** Find an expression for $\boldsymbol{\theta}$ that maximizes this expected complete-data log likelihood. You may find it helpful to use Lagrange multipliers in order to enforce the constraint $\sum \theta_k = 1$. Why does this optimal $\boldsymbol{\theta}$ make intuitive sense?
   - **Part 4.C** Find an expression for the $\{p_k\}_{k=1}^K$ that maximize the expected complete-data log likelihood. Why does this optimal $\{p_k\}_{k=1}^K$ make intuitive sense?

5. Suppose that this had been a classification problem. That is, you were provided the "true" components $\mathbf{z}_n$ for each observation x_n, and you were going to perform the classification by inverting the provided generative model (i.e. now you're predicting $\mathbf{z}_n$ given x_n). Could you reuse any of your derivations above to estimate the parameters of the model?

6. Finally, implement your solution (see `T5_P1.py` for starter code). You are responsible for implementing the `expected_loglikelihood`, `e_step` and `m_step` functions. Test it out with data given 10 samples from 3 components with $p_1 = .1$, $p_2 = .5$, and $p_3 = .9$. How does it perform? What if you increase the number of samples to 1000 from each of the components? What if you change $p_2 = .2$? Hypothesize reasons for the differences in performance when you make these changes. You may need to record five to ten trials (random restarts) in order to observe meaningful insights.

Solution

1. First expand the data likelihood to include the necessary sums over observations x_n.

$$\ln p(\{x_n\}_{n=1}^N; \boldsymbol{\theta}, \{p_k\}_{k=1}^K) = \ln \prod_{n=1}^N p(x_n; \boldsymbol{\theta}, \{p_k\}_{k=1}^K) = \sum_{n=1}^N \ln p(x_n; \boldsymbol{\theta}, \{p_k\}_{k=1}^K)$$

Next, consider computing $p(x_n; \boldsymbol{\theta}, \{p_k\}_{k=1}^K)$, which will allow us to marginalize out the latents $\mathbf{z}_n$.

$$p(x_n; \boldsymbol{\theta}, \{p_k\}_{k=1}^K) = \sum_{k=1}^K p(x_n, \mathbf{z}_n; \theta_k, p_k)$$

$$p(x_n; \boldsymbol{\theta}, \{p_k\}_{k=1}^K) = \sum_{k=1}^K p(\mathbf{z}_n; \theta_k) p(x_n | \mathbf{z}_n; \theta_k, p_k)$$

$$p(x_n; \boldsymbol{\theta}, \{p_k\}_{k=1}^K) = \sum_{k=1}^K \theta_k p(x_n | \mathbf{z}_n; \theta_k, p_k)$$

$$p(x_n; \boldsymbol{\theta}, \{p_k\}_{k=1}^K) = \sum_{k=1}^K \theta_k (1-p_k)^{x_n - 1} p_k$$

Which we can substitute back into our expression for the data likelihood to write it as

$$\boxed{\ln p(\{x_n\}_{n=1}^N; \boldsymbol{\theta}, \{p_k\}_{k=1}^K) = \sum_{n=1}^N \ln \Big(\sum_{k=1}^K \theta_k (1-p_k)^{x_n - 1} p_k \Big)}$$

Optimizing the likelihood is directly intractable, because the data likelihood is expressed in terms of the log of a sum, which does not have an analytical solution. There is no possible way to simplify the likelihood so as to remove the sum from the logarithmic function, which would make calculating the derivative of the likelihood intractable.

2. First expand the complete-data negative log likelihood to include the necessary sums over observations $(x_n, \mathbf{z}_n)$.

$$-\ln p(D; \boldsymbol{\theta}, \{p_k\}_{k=1}^K) = -\ln p(\{(x_n, \mathbf{z}_n)\}_{n=1}^N; \boldsymbol{\theta}, \{p_k\}_{k=1}^K)$$

$$-\ln p(D; \boldsymbol{\theta}, \{p_k\}_{k=1}^K) = -\sum_{n=1}^N \ln p(x_n, \mathbf{z}_n; \boldsymbol{\theta}, \{p_k\}_{k=1}^K)$$

Next, consider computing $p(x_n, \mathbf{z}_n; \boldsymbol{\theta}, \{p_k\}_{k=1}^K)$.

$$-\ln p(D; \boldsymbol{\theta}, \{p_k\}_{k=1}^K) = -\sum_{n=1}^N \ln \left[p(\mathbf{z}_n; \boldsymbol{\theta}, \{p_k\}_{k=1}^K) p(x_n | \mathbf{z}_n; \boldsymbol{\theta}, \{p_k\}_{k=1}^K) \right]$$

$$-\ln p(D; \boldsymbol{\theta}, \{p_k\}_{k=1}^K) = -\sum_{n=1}^N \Big[\ln p(\mathbf{z}_n; \boldsymbol{\theta}, \{p_k\}_{k=1}^K) + \ln p(x_n | \mathbf{z}_n; \boldsymbol{\theta}, \{p_k\}_{k=1}^K) \Big]$$

Apply the "power trick" to rewrite $\ln p(\mathbf{z}_n; \boldsymbol{\theta}, \{p_k\}_{k=1}^K)$.

$$-\ln p(D; \boldsymbol{\theta}, \{p_k\}_{k=1}^K) = -\sum_{n=1}^N \Big[\ln \big(\prod_{k=1}^K \theta_k^{z_{nk}} \big) + \ln p(x_n | \mathbf{z}_n; \boldsymbol{\theta}, \{p_k\}_{k=1}^K) \Big]$$

$$-\ln p(D;\boldsymbol{\theta},\{p_k\}_{k=1}^K) = -\sum_{n=1}^{N}\Big[\sum_{k=1}^{K} z_{nk}\ln\theta_k + \ln p(x_n|\mathbf{z}_n;\boldsymbol{\theta},\{p_k\}_{k=1}^K)\Big]$$

Next, we know that $p(x_n|\mathbf{z}_n;\boldsymbol{\theta},\{p_k\}_{k=1}^K) = \text{Geometric}(x_n;p_k)^{\mathbf{z}_n}$, so the complete-data negative log-likelihood is rewritten as.

$$-\ln p(D;\boldsymbol{\theta},\{p_k\}_{k=1}^K) = -\sum_{n=1}^{N}\Big[\sum_{k=1}^{K} z_{nk}\ln\theta_k + \ln[(1-p_k)^{x_n-1}p_k]^{\mathbf{z}_n}\Big]$$

$$-\ln p(D;\boldsymbol{\theta},\{p_k\}_{k=1}^K) = -\sum_{n=1}^{N}\Big[\sum_{k=1}^{K} z_{nk}\ln\theta_k + \mathbf{z}_n\ln[(1-p_k)^{x_n-1}p_k]\Big]$$

$\mathbf{z}_n\ln[(1-p_k)^{x_n-1}p_k]$ can be rewritten by marginalizing latents $\mathbf{z}_n$.

$$-\ln p(D;\boldsymbol{\theta},\{p_k\}_{k=1}^K) = -\sum_{n=1}^{N}\Big[\sum_{k=1}^{K} z_{nk}\ln\theta_k + \sum_{k=1}^{K} z_{nk}\ln[(1-p_k)^{x_n-1}p_k]\Big]$$

$$\boxed{-\ln p(D;\boldsymbol{\theta},\{p_k\}_{k=1}^K) = -\sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\Big[\ln\theta_k + \ln[(1-p_k)^{x_n-1}p_k]\Big]}$$

3. • **Part 3.A**
To find an expression for $\mathbf{q}_n$, we seek to find an expression for $q_{nk} = p(\mathbf{z}_n = \mathbf{C}_k|x_n;\boldsymbol{\theta},\{p_k\}_{k=1}^K)$, which can be rewritten using Bayes' Rule.

$$p(\mathbf{z}_n = \mathbf{C}_k|x_n;\boldsymbol{\theta},\{p_k\}_{k=1}^K) \propto p(\mathbf{z}_n = \mathbf{C}_k;\boldsymbol{\theta},\{p_k\}_{k=1}^K)p(x_n|\mathbf{z}_n = \mathbf{C}_k;\boldsymbol{\theta},\{p_k\}_{k=1}^K)$$

$$p(\mathbf{z}_n = \mathbf{C}_k|x_n;\boldsymbol{\theta},\{p_k\}_{k=1}^K) \propto \theta_k p(x_n|\mathbf{z}_n = \mathbf{C}_k;\boldsymbol{\theta},\{p_k\}_{k=1}^K)$$

$$p(\mathbf{z}_n = \mathbf{C}_k|x_n;\boldsymbol{\theta},\{p_k\}_{k=1}^K) \propto \theta_k(1-p_k)^{x_n-1}p_k$$

$$q_{nk} \propto \theta_k(1-p_k)^{x_n-1}p_k$$

This expression for q_{nk} is only proportional to the actual value of q_{nk}. To ensure that the k entries of $\mathbf{q}_n$ sum to 1, we must normalize our expression for q_{nk}.

$$\boxed{q_{nk} = \frac{\theta_k(1-p_k)^{x_n-1}p_k}{\sum_{k=1}^K \theta_k(1-p_k)^{x_n-1}p_k}}$$

Such that the expression for $\mathbf{q}_n$ is:

$$\mathbf{q}_n = \begin{bmatrix} \frac{\theta_1(1-p_1)^{x_n-1}p_1}{\sum_{k=1}^K \theta_k(1-p_k)^{x_n-1}p_k} \\ \cdots \\ \cdots \\ \frac{\theta_K(1-p_K)^{x_n-1}p_K}{\sum_{k=1}^K \theta_k(1-p_k)^{x_n-1}p_k} \end{bmatrix}$$

• **Part 3.B**
For each example x_n, we can easily calculate $\mathbf{q}_n$ when we're given parameters $\boldsymbol{\theta}$ and $\{p_k\}_{k=1}^K$. The first step would be to do K calculations, where we calculate $\theta_k(1-p_k)^{x_n-1}p_k$ for each k in K. We can make these calculations since we are given $\boldsymbol{\theta}$, $\{p_k\}_{k=1}^K$, and x_n. Next we find the sum of these K calculations (so that we've calculated $\sum_{k=1}^K \theta_k(1-p_k)^{x_n-1}p_k$). The final step is to divide each of the K calculations of $\theta_k(1-p_k)^{x_n-1}p_k$, so that for each component k (each k in K) we have values of $q_{nk} = \frac{\theta_k(1-p_k)^{x_n-1}p_k}{\sum_{k=1}^K \theta_k(1-p_k)^{x_n-1}p_k}$. Combining these entries (in order) into a $K \times 1$ vector will give us the calculation for $\mathbf{q}_n$.

4. • **Part 4.A**
The expected complete-data log likelihood is expressed as.

$$\mathrm{E}_{\mathbf{z}}\Big[\ln p(\{(x_n,\mathbf{z}_n)\}_{n=1}^N;\boldsymbol{\theta},\{p_k\}_{k=1}^K\Big]$$

Which can be expressed as a sum over the data points as.

$$\mathrm{E}_{\mathbf{z}}\Big[\ln p(\{(x_n,\mathbf{z}_n)\}_{n=1}^N;\boldsymbol{\theta},\{p_k\}_{k=1}^K\Big]=\sum_{n=1}^N \mathrm{E}_{\mathbf{z}_n}\Big[\ln p(x_n,\mathbf{z}_n,\boldsymbol{\theta},\{p_k\}_{k=1}^K)\Big]$$

Since the components are discrete, the expectation can be expressed as a sum over the components using the expectation estimates $\mathbf{q}_n$.

$$\mathrm{E}_{\mathbf{z}}\Big[\ln p(\{(x_n,\mathbf{z}_n)\}_{n=1}^N;\boldsymbol{\theta},\{p_k\}_{k=1}^K\Big]=\sum_{n=1}^N\sum_{k=1}^K q_{nk}\ln p(x_n,\mathbf{z}_n;\theta_k,p_k)$$

$$\mathrm{E}_{\mathbf{z}}\Big[\ln p(\{(x_n,\mathbf{z}_n)\}_{n=1}^N;\boldsymbol{\theta},\{p_k\}_{k=1}^K\Big]=\sum_{n=1}^N\sum_{k=1}^K q_{nk}\ln\big(p(x_n|\mathbf{z}_n;\theta_k,p_k)p(\mathbf{z}_n;\theta_k,p_k)\big)$$

$$\mathrm{E}_{\mathbf{z}}\Big[\ln p(\{(x_n,\mathbf{z}_n)\}_{n=1}^N;\boldsymbol{\theta},\{p_k\}_{k=1}^K\Big]=\sum_{n=1}^N\sum_{k=1}^K q_{nk}\big[\ln p(x_n|\mathbf{z}_n;\theta_k,p_k)+\ln p(\mathbf{z}_n;\theta_k,p_k)\big]$$

$$\mathrm{E}_{\mathbf{z}}\Big[\ln p(\{(x_n,\mathbf{z}_n)\}_{n=1}^N;\boldsymbol{\theta},\{p_k\}_{k=1}^K\Big]=\sum_{n=1}^N\sum_{k=1}^K q_{nk}\big[\ln\big((1-p_k)^{x_n-1}p_k\big)+\ln\theta_k\big]$$

$$\boxed{\mathrm{E}_{\mathbf{z}}\Big[\ln p(\{(x_n,\mathbf{z}_n)\}_{n=1}^N;\boldsymbol{\theta},\{p_k\}_{k=1}^K\Big]=\sum_{n=1}^N\sum_{k=1}^K q_{nk}\big[(x_n-1)\ln(1-p_k)+\ln p_k+\ln\theta_k\big]}$$

• **Part 4.B**
From Part 4.A, the expected complete-data log likelihood is

$$\mathrm{E}_{\mathbf{z}}\Big[\ln p(\{(x_n,\mathbf{z}_n)\}_{n=1}^N;\boldsymbol{\theta},\{p_k\}_{k=1}^K\Big]=\sum_{n=1}^N\sum_{k=1}^K q_{nk}\big[(x_n-1)\ln(1-p_k)+\ln p_k+\ln\theta_k\big]$$

which is subject to the constraint $\sum_{k=1}^K\theta_k=1$, or equivalently $\sum_{k=1}^K\theta_k-1=0$. Since we are seeking to optimize the expected complete-data log likelihood subject to this constraint, we can express our optimization problem using the Lagrangian multiplier λ:

$$L(\theta_k,\lambda)=\sum_{n=1}^N\sum_{k=1}^K q_{nk}\big[(x_n-1)\ln(1-p_k)+\ln p_k+\ln\theta_k\big]+\lambda(\sum_{k=1}^K\theta_k-1)$$

First, take the partial with respect to λ to find the first order optimality condition:

$$\frac{\partial}{\partial\lambda}L(\theta_k,\lambda)=\sum_{k=1}^K\theta_k-1=0\rightarrow\sum_{k=1}^K\theta_k=1$$

Next, take the partial with respect to θ_k to find the first order optimality condition:

$$\frac{\partial}{\partial\theta_k}L(\theta_k,\lambda)=\sum_{n=1}^N\frac{q_{nk}}{\theta_k}+\lambda=0$$

$$\frac{1}{\theta_k}\sum_{n=1}^{N} q_{nk} = -\lambda$$

$$\frac{1}{\theta_k} = -\frac{\lambda}{\sum_{n=1}^{N} q_{nk}}$$

$$\theta_k = -\frac{\sum_{n=1}^{N} q_{nk}}{\lambda}$$

Substitute the above equation into the first order optimality condition obtained by the partial with respect to λ.

$$\sum_{k=1}^{K} \theta_k = 1 \rightarrow \sum_{k=1}^{K} -\frac{\sum_{n=1}^{N} q_{nk}}{\lambda} = 1$$

$$-\sum_{k=1}^{K}\sum_{n=1}^{N} q_{nk} = \lambda$$

Now substitute the above expression for λ into the expression for θ_k to find the optimal expression for θ_k.

$$\theta_k = -\frac{\sum_{n=1}^{N} q_{nk}}{\lambda} \rightarrow \theta_k = \frac{\sum_{n=1}^{N} q_{nk}}{\sum_{k=1}^{K}\sum_{n=1}^{N} q_{nk}}$$

Observe that $\sum_{k=1}^{K}\sum_{n=1}^{N} q_{nk} = N$ since $\sum_{k=1}^{K} q_{nk} = 1$ and $\sum_{n=1}^{N} 1 = N$, therefore the optimal expression for θ_k is.

$$\boxed{\theta_k = \frac{\sum_{n=1}^{N} q_{nk}}{N}}$$

Such that the optimal expression for $\boldsymbol{\theta}$ is:

$$\boldsymbol{\theta} = \begin{bmatrix} \frac{\sum_{n=1}^{N} q_{n1}}{N} \\ \cdots \\ \cdots \\ \frac{\sum_{n=1}^{N} q_{nK}}{N} \end{bmatrix}$$

This expression for the optimal $\boldsymbol{\theta}$ makes intuitive sense since each θ_k is represented by the sample mean of q_{nk}, which corresponds to the estimated probability of component k.

- **Part 4.C**

From Part 4.A, the expected complete-data log likelihood is8

$$\mathrm{E}_{\mathbf{z}}\Big[\ln p(\{(x_n, \mathbf{z}_n)\}_{n=1}^{N}; \boldsymbol{\theta}, \{p_k\}_{k=1}^{K}\Big] = \sum_{n=1}^{N}\sum_{k=1}^{K} q_{nk}\big[(x_n - 1)\ln(1 - p_k) + \ln p_k + \ln\theta_k\big]$$

Deriving with respect to p_k to find the first order optimality condition:

$$\frac{\partial}{\partial p_k}\mathrm{E}_{\mathbf{z}}\Big[\ln p(\{(x_n, \mathbf{z}_n)\}_{n=1}^{N}; \boldsymbol{\theta}, \{p_k\}_{k=1}^{K}\Big] = \sum_{n=1}^{N} q_{nk}\big[\frac{x_n - 1}{1 - p_k} + \frac{1}{p_k}\big] = 0$$

$$\sum_{n=1}^{N} q_{nk}\frac{x_n - 1}{1 - p_k} + \sum_{n=1}^{N} q_{nk}\frac{1}{p_k} = 0$$

$$\frac{1}{p_k}\sum_{n=1}^{N} q_{nk} = -\frac{1}{1 - p_k}\sum_{n=1}^{N} q_{nk}(x_n - 1)$$

$$\frac{1-p_k}{p_k} = -\frac{\sum_{n=1}^{N} q_{nk}(x_n - 1)}{\sum_{n=1}^{N} q_{nk}}$$

$$\frac{1}{p_k} - 1 = -\frac{\sum_{n=1}^{N} q_{nk}(1 - x_n)}{\sum_{n=1}^{N} q_{nk}}$$

$$\frac{1}{p_k} = 1 - \frac{\sum_{n=1}^{N} q_{nk}(1 - x_n)}{\sum_{n=1}^{N} q_{nk}}$$

$$\frac{1}{p_k} = \frac{\sum_{n=1}^{N} q_{nk} - \sum_{n=1}^{N} q_{nk}(1 - x_n)}{\sum_{n=1}^{N} q_{nk}} = \frac{\sum_{n=1}^{N} q_{nk} - q_{nk}(1 - x_n)}{\sum_{n=1}^{N} q_{nk}}$$

$$\frac{1}{p_k} = \frac{\sum_{n=1}^{N} q_{nk} x_n}{\sum_{n=1}^{N} q_{nk}}$$

$$\boxed{p_k = \frac{\sum_{n=1}^{N} q_{nk}}{\sum_{n=1}^{N} q_{nk} x_n}}$$

The mean of the Geometric(p_k) is $1/p_k$. This expression for the optimal p_k makes intuitive sense because it is the inverse of the estimated empirical mean for component k, which is consistent with the mean of a geometric distribution.

5. Since we are given the "true" components of $\mathbf{z}_n$ of each observations x_n, we can start using the initial form of the complete-data log likelihood. First expand the complete-data log likelihood to include the necessary sums over observations $(x_n, \mathbf{z}_n)$.

$$\ln p(D; \boldsymbol{\theta}, \{p_k\}_{k=1}^K) = \ln p(\{(x_n, \mathbf{z}_n)\}_{n=1}^N; \boldsymbol{\theta}, \{p_k\}_{k=1}^K)$$

$$\ln p(D; \boldsymbol{\theta}, \{p_k\}_{k=1}^K) = \sum_{n=1}^{N} \ln p(x_n, \mathbf{z}_n; \boldsymbol{\theta}, \{p_k\}_{k=1}^K)$$

Since we are inverting the provided generative model, we split our joint probability into the marginal probability of x_n and the conditional probability $\mathbf{z}_n | x_n$.

$$\ln p(D; \boldsymbol{\theta}, \{p_k\}_{k=1}^K) = \sum_{n=1}^{N} \ln \left[p(x_n; \boldsymbol{\theta}, \{p_k\}_{k=1}^K) p(\mathbf{z}_n | x_n; \boldsymbol{\theta}, \{p_k\}_{k=1}^K) \right]$$

$$\ln p(D; \boldsymbol{\theta}, \{p_k\}_{k=1}^K) = \sum_{n=1}^{N} \left[\ln p(x_n; \boldsymbol{\theta}, \{p_k\}_{k=1}^K) + \ln p(\mathbf{z}_n | x_n; \boldsymbol{\theta}, \{p_k\}_{k=1}^K) \right]$$

Once again, we can employ the power trick to rewrite $\ln p(\mathbf{z}_n | x_n; \boldsymbol{\theta}, \{p_k\}_{k=1}^K)$ as:

$$\ln p(D; \boldsymbol{\theta}, \{p_k\}_{k=1}^K) = \sum_{n=1}^{N} \left[\ln p(x_n; \boldsymbol{\theta}, \{p_k\}_{k=1}^K) + \sum_{k=1}^{K} z_{nk} \ln \theta_k \right]$$

We can apply the PMF for x_n to rewrite $\ln p(x_n; \boldsymbol{\theta}, \{p_k\}_{k=1}^K)$. Next, we know that $p(x_n; \boldsymbol{\theta}, \{p_k\}_{k=1}^K) = \text{Geometric}(x_n; p_k)^{\mathbf{z}_n}$, so the log-likelihood is rewritten as.

$$\ln p(D; \boldsymbol{\theta}, \{p_k\}_{k=1}^K) = \sum_{n=1}^{N} \left[\sum_{k=1}^{K} z_{nk} \ln[(1-p_k)^{x_n - 1} p_k] + \sum_{k=1}^{K} z_{nk} \ln \theta_k \right]$$

$$\ln p(D; \boldsymbol{\theta}, \{p_k\}_{k=1}^K) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \Big[\ln[(1-p_k)^{x_n - 1} p_k] + \ln \theta_k \Big]$$

$$\ln p(D; \boldsymbol{\theta}, \{p_k\}_{k=1}^K) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \Big[(x_n - 1) \ln(1-p_k) + \ln p_k + \ln \theta_k \Big]$$

Observe that this log-likelihood appears identical to the expected complete-data log likelihood from part 4.A, where the only difference is that rather than using the q_{nk} estimates, this log-likelihood uses the indicator variables z_{nk}. Suppose that $\mathbf{q}_n$ were hard assignments rather than soft assignments, such that all of the entries of $\mathbf{q}_n$ are 0, except for one which is 1. This setting would be identical to the use of the one-hot vector $\mathbf{z}_n$. Therefore, we can reuse the derivations from part 4.B and part 4.C by substituting z_{nk} for q_{nk} in our optimal expressions for parameters θ_k and p_k. I will redo those derivations here as proof of this.

Finding the optimal θ_k: the log-likelihood expressed above is subject to the constraint $\sum_{k=1}^K \theta_k = 1$, or equivalently $\sum_{k=1}^K \theta_k - 1 = 0$. Since we are seeking to optimize the expected complete-data log likelihood subject to this constraint, we can express our optimization problem using the Lagrangian multiplier λ:

$$L(\theta_k, \lambda) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \big[(x_n - 1) \ln(1-p_k) + \ln p_k + \ln \theta_k \big] + \lambda (\sum_{k=1}^{K} \theta_k - 1)$$

First, take the partial with respect to λ to find the first order optimality condition:

$$\frac{\partial}{\partial \lambda} L(\theta_k, \lambda) = \sum_{k=1}^{K} \theta_k - 1 = 0 \rightarrow \sum_{k=1}^{K} \theta_k = 1$$

Next, take the partial with respect to θ_k to find the first order optimality condition:

$$\frac{\partial}{\partial \theta_k} L(\theta_k, \lambda) = \sum_{n=1}^{N} \frac{z_{nk}}{\theta_k} + \lambda = 0$$

$$\frac{1}{\theta_k} \sum_{n=1}^{N} z_{nk} = -\lambda$$

$$\frac{1}{\theta_k} = -\frac{\lambda}{\sum_{n=1}^{N} z_{nk}}$$

$$\theta_k = -\frac{\sum_{n=1}^{N} z_{nk}}{\lambda}$$

Substitute the above equation into the first order optimality condition obtained by the partial with respect to λ.

$$\sum_{k=1}^{K} \theta_k = 1 \rightarrow \sum_{k=1}^{K} -\frac{\sum_{n=1}^{N} z_{nk}}{\lambda} = 1$$

$$-\sum_{k=1}^{K} \sum_{n=1}^{N} z_{nk} = \lambda$$

Now substitute the above expression for λ into the expression for θ_k to find the optimal expression for θ_k.

$$\theta_k = -\frac{\sum_{n=1}^{N} z_{nk}}{\lambda} \rightarrow \theta_k = \frac{\sum_{n=1}^{N} z_{nk}}{\sum_{k=1}^{K} \sum_{n=1}^{N} z_{nk}}$$

Observe that $\sum_{k=1}^{K}\sum_{n=1}^{N} z_{nk} = N$ since $\sum_{k=1}^{K} z_{nk} = 1$ and $\sum_{n=1}^{N} 1 = N$, therefore the optimal expression for θ_k is.

$$\boxed{\theta_k = \frac{\sum_{n=1}^{N} z_{nk}}{N}}$$

Finding the optimal p_k: Derive with respect to p_k to find the first order optimality condition:

$$\frac{\partial}{\partial p_k} \ln p(D; \boldsymbol{\theta}, \{p_k\}_{k=1}^{K}) = \sum_{n=1}^{N} z_{nk} \big[\frac{x_n - 1}{1 - p_k} + \frac{1}{p_k}\big] = 0$$

$$\sum_{n=1}^{N} z_{nk} \frac{x_n - 1}{1 - p_k} + \sum_{n=1}^{N} z_{nk} \frac{1}{p_k} = 0$$

$$\frac{1}{p_k} \sum_{n=1}^{N} z_{nk} = -\frac{1}{1 - p_k} \sum_{n=1}^{N} z_{nk}(x_n - 1)$$

$$\frac{1 - p_k}{p_k} = -\frac{\sum_{n=1}^{N} z_{nk}(x_n - 1)}{\sum_{n=1}^{N} z_{nk}}$$

$$\frac{1}{p_k} - 1 = -\frac{\sum_{n=1}^{N} z_{nk}(1 - x_n)}{\sum_{n=1}^{N} z_{nk}}$$

$$\frac{1}{p_k} = 1 - \frac{\sum_{n=1}^{N} z_{nk}(1 - x_n)}{\sum_{n=1}^{N} z_{nk}}$$

$$\frac{1}{p_k} = \frac{\sum_{n=1}^{N} z_{nk} - \sum_{n=1}^{N} z_{nk}(1 - x_n)}{\sum_{n=1}^{N} z_{nk}} = \frac{\sum_{n=1}^{N} z_{nk} - z_{nk}(1 - x_n)}{\sum_{n=1}^{N} z_{nk}}$$

$$\frac{1}{p_k} = \frac{\sum_{n=1}^{N} z_{nk} x_n}{\sum_{n=1}^{N} z_{nk}}$$

$$\boxed{p_k = \frac{\sum_{n=1}^{N} z_{nk}}{\sum_{n=1}^{N} z_{nk} x_n}}$$

6. After running multiple trials, it was found that there is significant variation in the final results. Therefore, for each model 10 trials were run and the results were averaged over the 10 trials. This was done for four different models, where the results are shown in the table below:

Model	Average Final Probs	Average Accuracy
10 samples, $p_2 = .5$	[0.12757834 0.49482668 0.85287181]	0.6066666666666666
10 samples, $p_2 = .2$	[0.10778381 0.49737286 0.86552953]	0.6033333333333333
1000 samples, $p_2 = .5$	[0.10083284 0.47620877 0.86846178]	0.6384666666666666
1000 samples, $p_2 = .2$	[0.09312018 0.29478138 0.82875623]	0.6277333333333334

On average over 10 trials, the models that used 1000 samples had a better average accuracy than the models that used 10 samples. This makes intuitive sense, since if we provide the model with a larger sample of data from the geometric distribution, then we expect that the model will have a higher accuracy in estimating the geometric distribution. Moreover, increasing the sample size increased the number of iterations that the model went through since there are more data points for the model to fit to, which allows the model to better approximate the component probabilities. We find that reducing

the prior probability p_2 from $p_2 = .5$ to $p_2 = .2$ slightly lowers the average accuracy of our model for both the model with 10 samples and the model with 1000 samples. When $p_2 = .2$, the prior probabilities $p_1 = .1$ and $p_2 = .2$ are very close together and the model has a difficult time distinguishing between the two clusters when it is optimally fitting the data, which will adversely affect the clustering results (estimated assignments to components) and therefore reduce the accuracy.